

NPS-62GH0071A

# United States Naval Postgraduate School



AN EVALUATION OF FITNESS REPORTS SCALES

by

Wm. H. Githens

and

Lynn A. Lacey

15 July 1970

This document has been approved for public release  
and sale; its distribution is unlimited.



NAVAL POSTGRADUATE SCHOOL  
Monterey, California

Rear Admiral R. W. McNitt, USN  
Superintendent

R. F. Rinehart  
Academic Dean

ABSTRACT: AN EVALUATION OF FITNESS REPORTS SCALES

A sample of convenience was obtained in which 15 officers completed (anonymously) fitness reports on each other. Fitness report scales were examined to determine their quality based on the statistical considerations of "discrimination" and "disagreement" index. It was found that there was a greater spread in scores (fitness marks) when a number of judges rate one individual than when an average judge rates a number of individuals. Generalizing from the study is prohibited by the size and nature of the sample. The study demonstrates a type of analysis that can be performed and the type of information that can be obtained by studies of this type. Replications of the study are recommended.

THE UNIVERSITY OF CHICAGO

THE UNIVERSITY OF CHICAGO

THE UNIVERSITY OF CHICAGO

THE UNIVERSITY OF CHICAGO

THE UNIVERSITY OF CHICAGO

## TABLE OF CONTENTS

	Page No.
Title Page with Abstract	1
Table of Contents	3
List of Tables	5
Problem and Background	7
Procedure	8
Results	10
Conclusions and Recommendations	13
Appendix	15
Distribution List	17
DD Form 1473	19

# Blank Page

## LIST OF TABLES

	Page No.
Table 1 - Population Characteristics	8
Table 2 - Computation of Discrimination Index and Disagreement Index	9
Table 3 - Disagreement Index and Discrimination Index of Each Fitness Report Scale	12





## Problem and Background

How can one evaluate fitness report scales? Ideally, in measuring the effectiveness of a scale, one would have some objective measurement of an officer's work performance to which fitness report scale marks would be related. The fitness report scales on which the marks more closely reflected the objective measurement of work performance would, of course, be the more desirable scales. But such an objective measurement of job performance does not exist. If it did, there would be no need for the fitness report scales which involve human judgment and all its human errors, since the objective measurement of job performance itself would serve any purpose for which the scales are used.

Since the objective measurement of work performance does not exist, other bases for evaluating the scales must be used. One basis used to evaluate scales is to make a judgment of their relevancy. This is perhaps the most important aspect of any scale. Someone has evidently judged that the scales on the fitness report are relevant to the performance of officers. Otherwise they would not have been included on the form.

Another basis for evaluating scales is "statistical." This report is concerned with the "statistical" evaluation of scales. Assuming qualitative differences between ratees actually exist, statistically good scales have the following two characteristics:

- a. "Discrimination" between individuals, so that individuals are not all rated at the same level.
- b. Inter-rater reliability, or very little "disagreement" among raters when they are judging the same behavior.

These two statistical characteristics can then be used to evaluate fitness report scales. The information needed to evaluate the "disagreement" characteristic is not usually available.

A retired Commanding Officer of a destroyer made available the information making evaluation on both characteristics possible. Aboard his destroyer he became intensely interested in the judgmental evaluation of his officers. He had each of his fifteen officers complete the then operational fitness report [NavPers 310 (Rev. 4-62) presented as the Appendix to this report] on each of the other officers. The reports were completed as usual except that the raters remained anonymous. A summary of the sample's characteristics is presented in Table 1.

TABLE 1  
Population Characteristics

Rank	Frequency	USN(R)	Frequency	Designator	Frequency
ENS	4	USN	10	1100	9
LTJG	8	USN(R)	5	1105	5
LT	0			3100	1
LCDR	1				
CDR	2				
CAPT	0				
Total	15		15		15

#### Procedure

Points were assigned to each of the rating scales of items 14 through 20 as indicated on the report form. Item 14 consists of 9-point scales, items 15 and 16 consist of 5-point scales, and item 20

consists of 7-point scales. Comparisons among scales were restricted to comparisons within these three sets of scales (9, 5, and 7-point lengths).

Index numbers were generated to reflect the two statistical characteristics of discrimination and disagreement. Table 2 shows the computation procedure for obtaining the "discrimination index" and "disagreement" index.

TABLE 2

Computation of Discrimination Index and Disagreement Index

Each scale was analyzed as follows:

RATEES							
		1.	2.	3.	4.	5.	$\sigma$ of Each Row
R	A	x	x	x	x	x	$\sigma_A$
A	B	x	x	x	x	x	$\sigma_B$
T	C	x	x	x	x	x	$\sigma_C$
E	D	x	x	x	x	x	$\sigma_D$
R	E	x	x	x	x	x	$\sigma_E$
S							
							$\bar{x}_{\sigma_{A-E}} = \text{Discrimination Index}$
$\sigma$ of each column		$\sigma_1$	$\sigma_2$	$\sigma_3$	$\sigma_4$	$\sigma_5$	
							$\bar{x}_{\sigma_{1-5}} = \text{Disagreement Index}$

Each scale was examined individually. "Discrimination" was first determined for each rater. Even though the raters were not identified on the reports, it was possible to group reports of the same rater by

matching certain miscellaneous characteristics of the reports. The standard deviation of the marks the rater assigned for each scale across ratees (in Table 2 -- the standard deviation of rows) was computed, a numerically high standard deviation indicating good discrimination. To obtain a discrimination index for each scale across raters, a simple average (mean) of these standard deviations was computed (in Table 2 -- the average of the row standard deviations).

An index of "disagreement" was also generated for each scale. The standard deviation of the marks the raters assigned to each ratee was computed (in Table 2 -- the standard deviation of columns). To obtain a "disagreement index" across all ratees, a simple average (mean) of these standard deviations was computed (in Table 2 -- the average of the column standard deviations). These disagreement indexes are influenced by both the relative ratings assigned by raters (i.e., agreement among raters in their relative ordering of ratees) and agreement among raters in the absolute level of their ratings -- the aspect usually influenced by leniency error.

### Results

The results of the analysis are shown in Table 3. Most of the scales that were high in "disagreement" were low in "discrimination" and vice versa. A lack in either low disagreement or high discrimination reduces the utility of the scale. Five of the scales were relatively favorable on both the disagreement and discrimination scales. They are:



- 14e Performance - As' ( ) Watch Officer
- 14f Performance - Technical Specialty ( )
- 20k Leadership - Personal Behavior
- 20l Leadership - Military Behavior
- 20m Leadership - Self-expression (oral)

They constitute the best of the scales as determined by this statistical analysis.

Three of the scales were relatively poor on both disagreement and discrimination. They are:

- 16c Foreign Duty
- 20a Leadership - Professional Knowledge
- 20b Leadership - Moral Courage

The most significant finding, however, is the similarity in level of "discrimination" and "disagreement" indexes. Ideally, judges would rate an individual on a scale with perfect agreement; and, assuming that individual differences exist on a scale, their ratings would reflect the true range of individual differences on that scale. Of the 26 scales on the fitness report, 17 scales have disagreement values that numerically exceed their discrimination values. This finding indicates that for these 17 scales, there is a greater spread in scores when a number of judges rate one individual than when an average judge rates a number of individuals. In other words, in this sample the raters disagree on individuals' ratings on a scale to a greater extent than average raters are able to discriminate among individuals on the scale.

TABLE 3

Disagreement Index and Discrimination Index  
of Each Fitness Report Scale

Item	Title	No. Scale Points	Disagreement Index <sup>1</sup>	Discrimination Index <sup>2</sup>
14	<u>Performance</u>			
(a)	Present Assignment	9	1.30	1.16
(b)	Shiphandling and Seamanship	9	1.14	1.11
(c)	Airmanship	9	----	----
(d)	Collateral Duties	9	1.21	.98
(e)	As _____ Watch Officer	9	1.19	1.16
(f)	Technical Specialty (_____)	9	1.03	1.15
(g)	Command Potential or Ability	9	1.36	1.33
(h)	Administrative & Management Ability	9	1.37	1.15
15	Overall Evaluation	5	.61	.72
16	<u>Desirability</u>			
(a)	Operational	5	.84	.81
(b)	Staff or Administrative	5	.82	.81
(c)	Foreign Duty	5	.84	.78
20	<u>Leadership</u>			
(a)	Professional Knowledge	7	.89	.80
(b)	Moral Courage	7	.89	.80
(c)	Loyalty	7	.79	.76
(d)	Force	7	.91	.89
(e)	Initiative	7	.89	.88
(f)	Industry	7	.87	.87
(g)	Imagination	7	.82	.78
(h)	Judgment	7	.84	.79
(i)	Reliability	7	.85	.80
(j)	Cooperation	7	.90	.98
(k)	Personal Behavior	7	.79	.85
(l)	Military Behavior	7	.82	.94
(m)	Self-expression (oral)	7	.84	.84
(n)	Self-expression (written)	7	.73	.73

Notes:

<sup>1</sup>Mean standard deviation of ratings on same subjects by different raters.

<sup>2</sup>Mean standard deviation of ratings for different subjects by same raters.

## Conclusions and Recommendations

For this specific sample, the "statistically desirable" qualities of each scale were determined and the five "statistically best" scales were identified. In general it was found that raters differed among themselves on the ratings they assigned, to the same degree that an average rater discriminated among ratees. Some implications of this general finding are that: (1) fitness marks should be interpreted as being highly dependent on the particular rater involved, and (2) there is a need for training of raters and/or better definition of scales so that inter-rater agreement would be increased. It is recognized that only one of the 15 raters in this study (the Commanding Officer) was a practiced rater. But since specific training in rating is not normally provided for officers who will be expected to complete fitness reports, the lack of experience in 14 of the 15 raters of this study may not have reduced the representativeness of this sample.

It would be expected that ratings by peers would differ somewhat from ratings by superiors or ratings by subordinates. This study combined all three varieties (no choice due to anonymity of raters) and this undoubtedly accounts for some of the non-reliability among raters. The accumulation of evidence from many such studies where raters could be identified would reveal the specific ways in which superiors, peers, and subordinates differ in their ratings. Statistical corrections could then be applied in order to obtain a better estimate of inter-rater reliability.

The sample size of this study was too small to permit justifiably generalizing from the results. This study provides, however, a

demonstration of the type of analysis that can be performed and the type of information that can be obtained by studies of this type. Replications of this study within small clusters of officers who are familiar with each others' job performance would permit the accumulation of information from which generalizations could reasonably be made.



## APPENDIX

NAVPERS 310W (REV. 4-62)

## REPORT ON THE FITNESS OF OFFICERS

## WORKSHEET

1. NAME (Last, first, middle)	2. GRADE	3. USN(R)	4. DESIGNATOR	5. FILE NUMBER
6. SHIP OR STATION			7. DATE REPORTED PRESENT DUTY STATION	
8. OCCASION FOR REPORT <input type="checkbox"/> PERIODIC <input type="checkbox"/> DETACHMENT OF REPORTING SENIOR <input type="checkbox"/> DETACHMENT OF OFFICER			9. TYPE OF REPORT <input type="checkbox"/> REGULAR <input type="checkbox"/> CON-CURRENT <input type="checkbox"/> SPECIAL	
10. PERIOD OF REPORT FROM: _____ TO: _____				
11. DUTIES (List principal duties assigned and the number of months during the period for which assigned)				

12. EMPLOYMENT OF COMMAND DURING PERIOD OF THIS REPORT

13. REFERENCE HERE AND APPEND ANY COMMENDABLE OR ADVERSE REPORTS ON THIS OFFICER RECEIVED DURING THE PERIOD OF THIS REPORT

14. PERFORMANCE OF DUTIES (Evaluate his performance of duty in comparison with other officers of his grade and approximate length of service)

DUTY ASSIGNMENT	NOT OBS. OR N.A.	Outstanding performance. (9) (8)	Excellent performance. Frequently demonstrates outstanding performance. (7) (6)	Very good performance. Frequently demonstrates excellent performance. (5) (4)	Satisfactory performance. Basically qualified. (3) (2)	Inadequate performance. He is not qualified. (1) (Adverse)
(a) PRESENT ASSIGNMENT						*
(b) SHIPHANDLING AND SEAMANSHIP						*
(c) AIRMANSHIP						*
(d) COLLATERAL DUTIES						*
(e) AS _____ WATCH OFFICER						*
(f) TECHNICAL SPECIALTY (_____)						*
(g) COMMAND POTENTIAL OR ABILITY						*
(h) ADMINISTRATIVE AND MANAGEMENT ABILITY						*

15. OVERALL EVALUATION: (a) In comparison with other officers of his grade and approximate length of service, how would you designate this officer?  
 (b) For this report period indicate in (b) how many officers of his grade you have designated in each category of (a).

	NOT OBSERVED	One of the highly outstanding officers I know (5)	A very fine officer of great value to the service (4)	A dependable and typically effective officer (3)	An acceptable officer (2)	Unsatisfactory (1) (Adverse)
(a)						*
(b)						

16. DESIRABILITY: Considering (1) the possible requirements of war and peace, (2) this officer's professional and technical competence, and (3) the adaptability of this officer to the varying conditions of naval service, indicate your attitude toward having this officer under your command in the following type of assignments: (1)

	NOT OBSERVED	Particularly desire (5)	Prefer to most (4)	Pleased to have (3)	Satisfied to have (2)	Prefer not to have (Adverse) (1)
(a) OPERATIONAL						*
(b) STAFF OR ADMINISTRATIVE						*
(c) FOREIGN DUTY						*

17. ENTRIES ON THIS REPORT ARE BASED ON (Check appropriate box)

☐ DAILY CONTACT AND CLOSE OBSERVATION    ☐ FREQUENT OBSERVATION    ☐ INFREQUENT OBSERVATION    ☐ RECORDS AND REPORTS ONLY

18. FOR FUTURE ASSIGNMENTS:

Based on your observations, for what type of duty do you consider him best qualified for his next assignment at sea and shore?

SEA \_\_\_\_\_

SHORE \_\_\_\_\_

Comment, if appropriate

19. NAME, GRADE, FILE NUMBER, DESIGNATOR AND OFFICIAL TITLE OF REPORTING SENIOR.

20. LEADERSHIP: In comparison with other officers of his grade and approximate length of duty assignment, to what degree has this officer exhibited the following qualities of leadership?

## DEFINITIONS

OUTSTANDING - ONE out of 100 - Exceeds ALL others  
 EXCEPTIONAL - One of the next top FEW - Extraordinary  
 SUPERIOR - ABOVE the great MAJORITY  
 EXCELLENT - EQUAL to the majority

ACCEPTABLE - BELOW the majority  
 MARGINAL - Barely satisfactory  
 UNSATISFACTORY

	NOT OBSERVED	OUT OF 100 OUTSTANDING	EXCEPTIONAL	SUPERIOR	EXCELLENT	ACCEPTABLE	MARGINAL	UNSATISFACTORY
(a) PROFESSIONAL KNOWLEDGE (Comprehension of all aspects of the profession)								*
(b) MORAL COURAGE (To do what he ought to do regardless of consequences to himself)								*
(c) LOYALTY (His faithfulness and allegiance to his shipmates, his command, the service and the nation)								*
(d) FORCE (The positive and enthusiastic manner with which he fulfills his responsibilities)								*
(e) INITIATIVE (His willingness to seek out and accept responsibility)								*
(f) INDUSTRY (The zeal exhibited and energy applied in the performance of his duties)								*
(g) IMAGINATION (Resourcefulness, creativeness, and capacity to plan constructively)								*
(h) JUDGMENT (His ability to develop correct and logical conclusions)								*
(i) RELIABILITY (The dependability and thoroughness exhibited in meeting responsibilities)								*
(j) COOPERATION (His ability and willingness to work in harmony with others)								*
(k) PERSONAL BEHAVIOR (His demeanor, disposition, sociability and sobriety)								*
(l) MILITARY BEARING (His military carriage, correctness of uniform, smartness of appearance and physical fitness)								*
(m) SELF-EXPRESSION (ORAL) (His ability to express himself orally)								*
(n) SELF-EXPRESSION (WRITTEN) (His ability to express himself in writing)								*

21. COMMENTS: (Reporting seniors are encouraged to discuss this report with the officer, but not necessarily show it.)

- (a) Make comments regarding any strengths, special accomplishments, contributions to the Naval and National service, or minor weaknesses. (Minor weaknesses must be discussed with the officer)

Have minor weaknesses been discussed with officer? .

☐ YES

☐ NO

☐ NOT APPLICABLE

- \*(b) ADVERSE COMMENTS, if any. Comments in this section are mandatory for adverse or unsatisfactory marks in section 14, 15, 16 and 20. Reports containing adverse matter must be referred for statement pursuant to Art. 1701.8, Navy Regulations. Statement of officer must be attached to this report. (Marks in starred (\*) boxes are adverse.)

Has officer seen this report?

☐ YES ☐ NO

- (c) What has been the trend of his performance since your last report?

☐ FIRST REPORT ☐ IMPROVING ☐ CONSISTENT ☐ DECLINING

22. DATE FORWARDED

SIGNATURE OF REPORTING SENIOR

23. CONCURRENT REPORT:

DATE FORWARDED

SIGNATURE OF REGULAR REPORTING SENIOR

# INITIAL DISTRIBUTION LIST

	No. of Copies
1. Defense Documentation Center Cameron Station Alexandria, Virginia 22314	20
2. Library Naval Postgraduate School Monterey, California 93940	2
3. Chief of Naval Personnel (Pers-A3) Bureau of Naval Personnel Department of the Navy Washington, D. C. 20370	25
4. Commanding Officer Naval Personnel Research and Development Laboratory Washington Navy Yard Washington, D. C. 20390	3
5. U. S. Army Personnel Research Office Office, Chief Research and Development Department of the Army Washington, D. C. 20315	3
6. Commandant U. S. Marine Corps (Code A01B) Washington, D. C. 20380	1
7. Commanding Officer Air Force Personnel Laboratory (ASD - AFSC) Lackland Air Force Base San Antonio, Texas 78558	1
8. Commanding Officer Naval Personnel & Training Research Laboratory San Diego, California 92152	30
9. Dr. William Githens (Code 62Gh) Naval Postgraduate School Monterey, California 93940	20
10. Mr. Lynn Lacey Naval Personnel & Training Research Laboratory San Diego, California 92152	10

No. of Copies

11. Dean of Research Administration  
Naval Postgraduate School  
Monterey, California 93940

2



## DOCUMENT CONTROL DATA - R &amp; D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)

Naval Postgraduate School  
Monterey, California 93940

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

An Evaluation of Fitness Reports Scales

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Technical Report, 1970

5. AUTHOR(S) (First name, middle initial, last name)

William H. Githens  
Lynn A. Lacey

6. REPORT DATE

15 July 1970

7a. TOTAL NO. OF PAGES

20

7b. NO. OF REFS

8a. CONTRACT OR GRANT NO.

b. PROJECT NO.

9a. ORIGINATOR'S REPORT NUMBER(S)

NPS-62GH0071A

c.

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

d.

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

Naval Postgraduate School  
Monterey, California 93940

13. ABSTRACT

A sample of convenience was obtained in which 15 officers completed (anonymously) fitness reports on each other. Fitness report scales were examined to determine their quality based on the statistical considerations of "discrimination" and "disagreement" index. It was found that there was a greater spread in scores (fitness marks) when a number of judges rate one individual than when an average judge rates a number of individuals. Generalizing from the study is prohibited by the size and nature of the sample. The study demonstrates a type of analysis that can be performed and the type of information that can be obtained by studies of this type. Replications of the study are recommended.

### KEY WORDS

LINK A

LINK B

LINK C

ROLE

WT

ROLE

WT

ROLE

WT

Fitness report

## Appraisal

## Statistical analysis

## Raters

## Performance

## Evaluation

## Reliability

5015108

DUDLEY KNOX LIBRARY - RESEARCH REPORTS



5 6853 01058150 7

